doi:10.1016/j.jmb.2004.06.058 J. Mol. Biol. (2004) 341, 1295–1315

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds

Douglas D. Axe*

The Babraham Institute
Structural Biology Unit
Babraham Research Campus
Cambridge CB2 4AT, UK

Proteins employ a wide variety of folds to perform their biological functions. How are these folds first acquired? An important step toward answering this is to obtain an estimate of the overall prevalence of sequences adopting functional folds. Since tertiary structure is needed for a typical enzyme active site to form, one way to obtain this estimate is to measure the prevalence of sequences supporting a working active site. Although the immense number of sequence combinations makes wholly random sampling unfeasible, two key simplifications may provide a solution. First, given the importance of hydrophobic interactions to protein folding, it seems likely that the sample space can be restricted to sequences carrying the hydropathic signature of a known fold. Second, because folds are stabilized by the cooperative action of many local interactions distributed throughout the structure, the overall problem of fold stabilization may be viewed reasonably as a collection of coupled local problems. This enables the difficulty of the whole problem to be assessed by assessing the difficulty of several smaller problems. Using these simplifications, the difficulty of specifying a working β-lactamase domain is assessed here. An alignment of homologous domain sequences is used to deduce the pattern of hydropathic constraints along chains that form the domain fold. Starting with a weakly functional sequence carrying this signature, clusters of ten side-chains within the fold are replaced randomly, within the boundaries of the signature, and tested for function. The prevalence of low-level function in four such experiments indicates that roughly one in 10^{64} signature-consistent sequences forms a working domain. Combined with the estimated prevalence of plausible hydropathic patterns (for any fold) and of relevant folds for particular functions, this implies the overall prevalence of sequences performing a specific function by any domain-sized fold may be as low as 1 in 10^{77}, adding to the body of evidence that functional folds require highly extraordinary sequences.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: functional constraints; sequence-function relationship; sequence-structure relationship; function landscape; sequence space

*Corresponding author

Introduction

Every quantifiable function that can be performed by proteins has a definite mapping onto the conceptual space representing all protein sequences. What can be discovered about these functional maps? Although the immense size of sequence space greatly limits the utility of direct experimental exploration, the sparse sampling that is feasible ought to be of use in addressing the most basic question of the overall prevalence of function. Progress on this front will both enhance our understanding of how new functional proteins arise naturally and inform our approach to generating them artificially.

This is a difficult problem to approach experimentally, however, and no clear picture has yet emerged. A number of studies have suggested that functional sequences are not extraordinarily rare,[1–5] while others have suggested that they are.[6–9] One of two approaches is typically used in these studies. The first, which could be termed the forward approach, involves producing a large collection of

Abbreviations used: MIC, minimum inhibitory concentration; indels, insertions and deletions.

E-mail address of the corresponding author: doug.axe@bbsrc.ac.uk

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.

sequences with no specified resemblance to known functional sequences and searching either for function or for properties generally associated with functional proteins. If the relevant sort of properties can be found among more or less random sequences, this provides a direct demonstration of their prevalence. The second approach works in reverse from an existing functional sequence. Here, the question is how much randomization a sequence known to have the relevant sort of function can withstand without losing that function.

Although both approaches have provided important insights, they may have drawbacks that contribute to the apparent discrepancies. The forward approach has not produced a sequence with properties that place it unequivocally among natural functional sequences. Whether the properties that have been found (e.g. proteolytic stability[10] or cooperative denaturation[1]) actually warrant such placement therefore remains an open question. On the other hand, because the reverse approach starts with a sequence that is not just functional but often nearly optimal, it may fail to take account of sequences having the relevant functional properties in a very rudimentary form. Also the difficulty of taking proper account of sequence context presents itself when natural proteins are studied by making one or a few substitutions at a time.[8] Substitutions found to be functionally tolerable in such experiments might be tolerable only because the vast majority of the protein remains untouched.[11]

In light of these difficulties, an important first step in the present study is to consider carefully what we mean by function in the first place. Different answers to this may well lead to different experimental approaches and different conclusions, each valid when properly understood. The focus here will be upon enzymatic function, by which we mean not mere catalytic activity but rather catalysis that is mechanistically enzyme-like, requiring an active site with definite geometry (at least during chemical conversion) by which particular side-chains make specific contributions to the overall catalytic process. The focus, then, will be on mode of catalysis rather than rate. The justification for this is that there is a clear connection between active-site formation and protein folding, in that active sites generally require the local positioning of multiple side-chains that are dispersed in the sequence. Something akin to tertiary structure, however crude, must therefore emerge in working form before natural selection can begin the process of refining a new fold. By assessing the difficulty of achieving the sort of structure needed to form a working active site, we therefore gain insight into a critical step in the emergence of new protein folds.

How might the other difficulties be avoided? A recent study of the requirements for chorismate mutase function *in vivo* demonstrates a promising approach.[9] Chorismate mutase gene libraries prepared in that work were constrained to preserve all active-site residues and the sequential arrangement of hydrophobic and hydrophilic side-chains present in a natural version of the enzyme. Within these constraints, though, specific residue assignments were essentially random, resulting in numerous disruptive changes throughout the encoded proteins. This is an example of the reverse approach, in that it uses a natural sequence as a starting point but, because the produced variants carry extensive disruption throughout the structure rather than just local disruption, they provide reliable information on the stringency of functional requirements. The

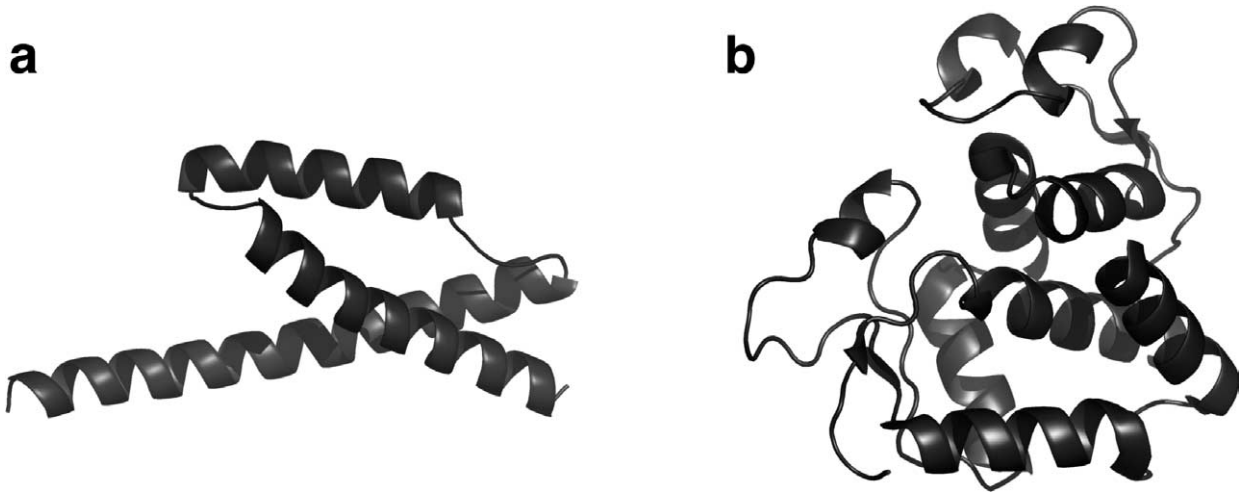

Figure 1. Relative fold complexities of the chorismate mutase monomer and the β-lactamase large domain. a, The AroQ-type chorismate mutase examined by Taylor *et al.*[9] is formed by symmetrical association of a pair of 93 residue monomers with this three helix structure (PDB entry 1ECM). b, The TEM-1 penicillinase, a typical class A β-lactamase, functions as a 263 residue monomer with two structural domains, the larger one shown here (153 residues; PDB entry 1ERM). This fold is made more complex by its larger size, and by the number of structural components (loops, helices, and strands) and the degree to which formation of these components is intrinsically coupled to the formation of tertiary structure (as is generally the case for strands and loops, but not for helices).

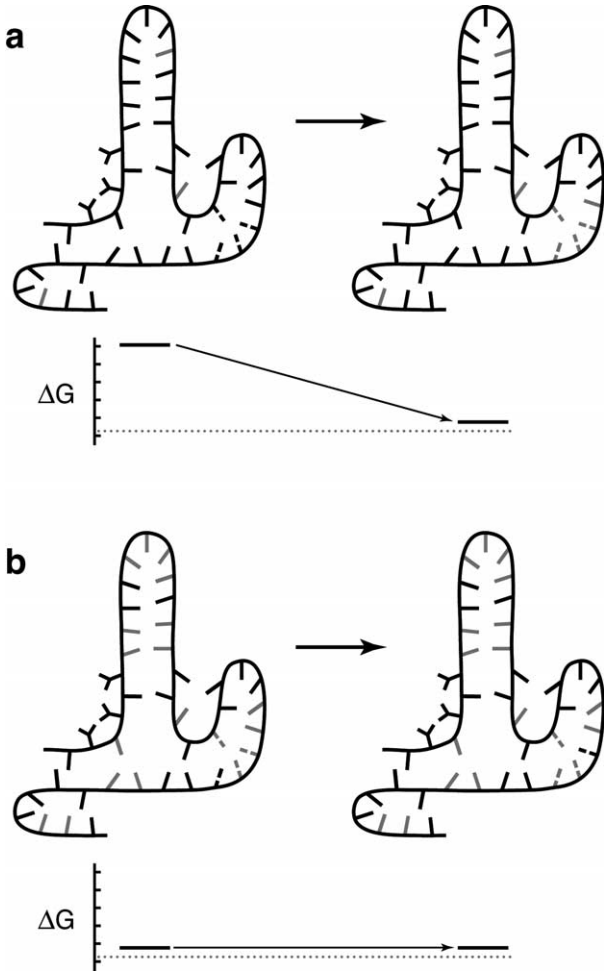

Figure 2. Importance of starting sequence and selection threshold on local side-chain randomization experiments. A generic enzyme is represented schematically as a backbone conformation (curved line) stabilized by a large number of interactions among side-chains (appendages) distributed throughout the structure, resulting in formation of a working active site (Y-shaped appendages). Each black appendage represents an "optimal" side-chain, meaning that it stabilizes the native fold at least as well as any of its 19 possible replacements would in the same context. Grey appendages represent side-chains that are in this sense suboptimal. A set of structurally local side-chains (broken lines) is chosen for randomization with subsequent functional selection. Folding stabilities of the starting sequence and a passing randomized sequence are represented by a qualitative graph, with the dotted line representing the minimum stability for passing under the chosen selection conditions. a, Natural selection ensures that a wild-type starting sequence (left) has relatively few suboptimal side-chains. (Substitutions that improve the stability of natural proteins are therefore relatively rare, as data collated by Guerois *et al*. bear out. See Figure 5 of Guerois *et al.*,[47] disregarding data from reverse mutations [cyan squares]. Consequently, anything but the most stringent selection will count randomized variants that are significantly less stable (right) as "active". b, A uniformly suboptimal starting sequence having just enough activity to pass a very low selection threshold (left) ensures that randomized variants passing that threshold (right) retain interactions within the randomized region that are comparable in quality to those of the starting sequence.

prevalence of functional chorismate mutases among sequences carrying the specified hydropathic pattern was estimated to be just one in 10^{24}.[9]

In view of the rarity of sequences carrying that pattern (among all possible sequences) and the relative simplicity of the chorismate mutase fold (Figure 1a), this result suggests that sequences encoding working enzymes may generally be very rare. Further exploration of this possibility should address two points. First, it is important that enzyme folds of more typical complexity be examined. And second, since many different folds might be comparably suited to any given enzymatic function, it is important that we have some way to factor this in. In other words, if the prevalence of sequences performing a particular function enzymatically is our primary interest, then our analysis must not presume the necessity of any particular fold.

Because protein structures show natural division into compact folding units, called domains,[12] it is appropriate to frame the problem at this level. Here, the larger of the two domains forming β-lactamases of the class A variety (henceforth, the large domain) is used as a model system for assessing the requirements for functional formation of a moderately complex fold (Figure 1b). Although predominantly composed of α-helices, this domain contains small sheet regions and significant loop structure which, along with its size (just over 150 amino acid residues), make its complexity more representative of known domain folds. Another typical feature of domains, the ability to form specific associations with other domains, is ensured by the location of the β-lactamase active-site cleft at the interface between the large and small domains. As in the chorismate mutase study, disruptive substitutions throughout the large domain will provide a marginally adequate sequence context in which to assess the requirements for low-level function. By making use of sequence information from numerous related β-lactamases, it is possible to frame the analysis of this single fold in such a way that it illuminates the key aspects of the sequence-function relationship that must be explored in order to assess the overall prevalence of enzymatic function.

Experimental Approach

The use of mixed-base oligonucleotides for simultaneous randomization of a complete sequence (as in the chorismate mutase work[9]) becomes increasingly problematic for longer sequences. An alternative approach, applicable to sequences of any length, is first to degrade the

Although these interactions are not optimal, they favour the folded structure to a degree that is characteristic of a marginally functional enzyme fold, which cannot be said of the randomized interactions of a (right).

whole fold by widespread substitution and then to produce libraries having locally randomized regions within this barely adequate initial structure. Sequence constraints may then be assessed by the frequency of functional variants in these libraries. The importance of having an extensively degraded initial sequence may be illustrated more fully by considering the effect of the selection threshold on the outcome.

Most studies using a biological screen or selection method to score variants of a natural sequence as active or inactive employ a threshold that requires only a small fraction of wild-type activity for an active score to be assigned.[11] Coupled with the fact that natural proteins are typically folded with stabilities well in excess of the bare minimum under the conditions of selection, this means that variants scored active may actually carry significant structural disruption. As an illustration, consider an experiment in which random substitutions are introduced into a small region within a natural enzyme, with functional selection applied in the usual way (Figure 2a). Because the wild-type protein is well stabilized by favourable side-chain interactions throughout the fold (Figure 2a, left), it has some capacity to absorb the destabilizing effects of disruptive substitutions in small numbers (Figure 2a, right). In essence, the relatively high quality of interactions throughout the unchanged portion of the protein can compensate for, or buffer, the effects of unfavourable interactions within the changed portion. This accounts for the observation that substitutions having little functional effect alone or in modest numbers have very substantial disruptive effects when combined in numbers large enough to exhaust that initial buffering capacity.[8]

The buffering effect is problematic for measurement of sequence constraints simply because side-chain interactions in the randomized region are apt to be much less favourable in variants isolated by selection than in the initial sequence. If we intend to assess constraints by assessing the proportion of randomized variants that pass selection, we must ensure that any significant deterioration upon randomization will prevent passing. So, to assess the minimal constraints for proper enzyme function, the approach should be first to obtain an extensively degraded reference sequence that just passes a low selection threshold (Figure 2b, left) and then to subject locally randomized variants of that sequence to selection at the same threshold (Figure 2b, right). Because the reference sequence has virtually no capacity to buffer the effects of further disruption, the quality of side-chain interactions within the randomized region must be maintained in order for a variant to pass. By performing several such experiments at various locations in the structure, it should therefore be possible to estimate the fraction of side-chain specifications providing interactions that are just sufficiently favourable to support low-level enzyme function.

One way to produce the reference sequence is to introduce numerous amino acid substitutions more or less randomly into a natural sequence. Because each substitution affects the modified side-chain and its interaction partners, the number of residues perturbed is considerably larger than the number of changes introduced. Yet, even though a sequence produced in this way will be degraded substantially, some residues or pockets of residues will probably remain optimal in the sense used in Figure 2 (i.e. the best side-chain for that position in that context). In particular, if some side-chains have pivotal roles in stabilizing the native fold, these will be preserved in the reference sequence.

Such pivotal residues must be considered in the design of the local randomization experiments. For technical reasons (explained below) it will not be feasible for local randomization to be performed at all amino acid positions in the reference sequence. The constraints for forming a functioning large domain will instead be sampled in four separate randomization experiments covering just over a quarter of the positions. The positions sampled will therefore need to be reasonably representative of the whole domain, and it is particularly important that pivotal residues not be over-represented if we want to avoid exaggerating the constraints.

Results and Discussion

Identification of lower-bound selection threshold

The natural function of β-lactamases, protecting bacteria from the effects of penicillin-like antibiotics, provides a simple means of selecting functional variants over a wide range of thresholds. As with any selection system, though, there are limits to the useful range. At the low end, *Escherichia coli* strains have some innate resistance to common penicillins as a result of both uninducible, low-level hydrolytic activity of AmpC and the action of the AcrAB multidrug efflux system.[13] By the usual index of resistance (minimum inhibitory concentration, abbreviated MIC), the *E. coli* strain used in this work has innate ampicillin resistance measuring 5 μg/ml, meaning that it fails to produce visible colonies at 25 °C when ampicillin is present at concentrations equalling or exceeding this (see Materials and Methods for details of standard test conditions).

In principle, then, we can select ampicillin-resistant clones without interference from innate resistance by using this level of antibiotic. However, when attempts were made to produce a reference sequence using this selection threshold, sequences that passed selection were found to carry mutations that would eliminate function by the known enzymatic mechanism. For example, a 36 residue deletion tolerated at this threshold precludes formation of much of the active-site cleft by removing a substantial part of the large-domain

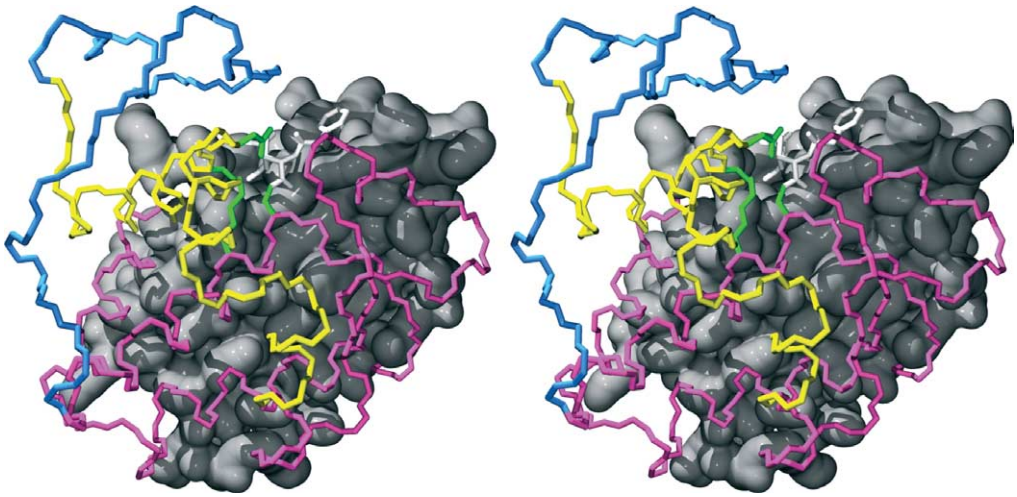

Figure 3. Structural importance of the 36 residue segment missing from a deletion mutant. The backbone structure is shown in stereo for the TEM-1 large-domain (PDB entry 1FQG) with space-filling representation of the small domain. The missing segment (yellow) includes two important active-site side-chains (Ser130 and Asn132, green). Two other active-site side-chains (Ser70 and Lys73, also green) are found not to be important for the low-level activity of the deletion mutant. Penicillin (white) is shown attached covalently *via* Ser70, representing the normal acyl-enzyme intermediate in the hydrolysis reaction.[14] As a consequence of the deletion, the blue portion of the chain cannot adopt its normal conformation.

core, eliminates two important catalytic residues, and prevents a stretch of 29 remaining residues from adopting its original conformation (Figure 3). Residues crucial to the function of class A β-lacta-lactamases (Ser70 and Lys73)[14] can be replaced in this deletion mutant without affecting its ability to confer resistance at this level. Whatever the mechanism of this resistance, then, it is safe to conclude on the basis of this evidence that it differs fundamentally from the well-studied mechanism of class A β-lactamases.[15] A reasonable conjecture, in view of the susceptibility of ampicillin to hydrolysis by simple acid or base catalysis,[16] is that polypeptides may promote ampicillin hydrolysis at low but detectable rates simply by displaying appropriately acidic or basic groups, in a manner analogous to peptide-catalyzed hydrolysis of RNA.[11,17]

Assessing the sequence constraints for this uncharacterized mechanism would be a worthwhile step toward characterizing it. A preliminary randomization experiment shows the constraints to be very low (unpublished results), consistent with the indifference to alteration described above. But in view of our present aim, assessment of the constraints entailed by a functional enzyme-like active site (see Introduction), we will need to exclude activities that do not meet this condition. The sequence carrying the 36 residue deletion is found to confer an ampicillin MIC of 10 μg/ml, which amounts to 0.1% of wild-type TEM-1 activity (TEM-1 MIC = 5200 μg/ml; (10–5)/(5200–5) = 0.001). If this is typical of sequences working by the uncharacterized mechanism, interference from such sequences will be eliminated by placing the selection threshold at this level.

Homologous sequence alignment

Both experimental stages of this study, production of the large-domain reference sequence and local randomization of that sequence, were guided by information present in an alignment of natural sequences that encode very similar domain folds. The SCOP structure classification (release 1.63†) lists 13 "species"-level variants of the class A β-lactamase fold. Removal of two of these (the TEM-52 variant being very similar to the TEM-1 variant, and the PER-1 variant showing substantial structural deviation from otherwise conserved features[18]) leaves a set of 11 natural large-domain variants with close structural similarity (Figure 4) and considerable sequence diversity.

This set can be enlarged to expand its diversity while maintaining tight structural similarity by including sequences with sufficient similarity to one of the structural representatives. Sequences having at least 50% side-chain identity typically have shared backbone structures encompassing 90% or more of their residues.[19] Using this as a cut-off, a search of the SwissProt database yields 33 additional natural domain sequences (after removal of virtual duplicates; see Materials and Methods). The resulting set of 44 homologues provides substantial sequence diversity, while permitting sequence alignment with very little ambiguity (Figures 5 and 6).

† http://scop.mrc-lmb.cam.ac.uk/scop

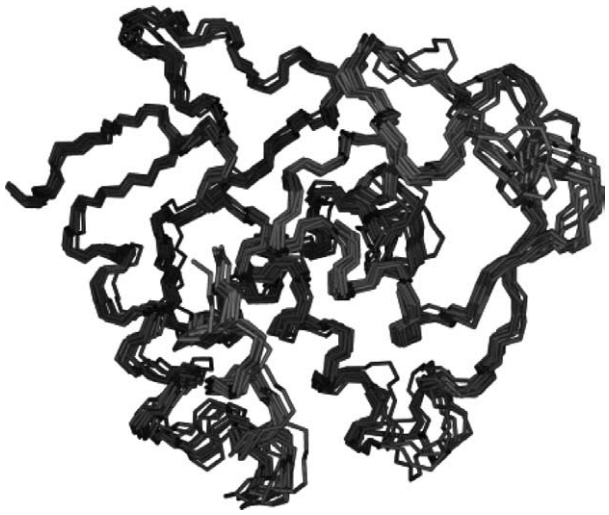

Figure 4. Superposition of the large-domain backbone structures of 11 class A β-lactamases. Structural data are from PDB entries: 1BSG, 1BUE, 1BZA, 1DY6, 1ERM, 1G6A, 1GHP, 1HZO, 1MFO, 1SHV, and 4BLM. Excluding hydrogen atoms, backbone RMS deviations from the TEM-1 structure (1ERM) are, in the above order: 0.82, 0.85, 0.85, 0.89, 0, 0.76, 1.24, 0.75, 0.86, 0.44, and 0.63 Å over alignments covering at least 87% of the full domain. The 1BSG and 1GHP structures show the largest RMS deviation (1.50 Å over an alignment covering 90% of the domain).

Finding a reference sequence

Dramatic loss of enzyme function can be achieved with a small number of highly disruptive changes, even without direct modification of the active site.[11] The objective here, however, is to introduce a large number of mildly disruptive changes so as to render many side-chains sub-optimal throughout the fold (Figure 2b, left). This is best achieved by introducing many changes together, without intervening selection. But in order for this not to cause complete disruption, it is necessary to mitigate somewhat the likely disruption at each position.

Using the wild-type TEM-1 sequence as a starting point, this was accomplished by limited substitution at five groups of positions (58 positions in total) across the large domain (see Materials and Methods). Substitution was limited in three respects. First, positions where side-chains form the active site were excluded from the groups chosen for change. Second, the wild-type TEM-1 residue was included as a possible alternative at 47 of the 58 positions, the remaining 11 positions having relatively uncommon residues in the TEM-1 sequence (Figure 6). And third, residue options were biased strongly toward side-chains represented in the alignment. In the first four substitution groups, 120 of the 122 possibilities allowed at the 49 affected positions are represented (Figure 6).

Substitutions in these first four groups were combined to produce a library of variants that had been subjected to limited substitution at 49 positions. At this point, ampicillin at the threshold level (10 μg/ml) was first used to select functional variants. Of several sequences found to permit growth, one with a better than average MIC (>40 μg/ml) was chosen as the progenitor of the reference sequence. The final step in producing the reference sequence coincided with the first local side-chain randomization experiment, as described below. After this randomization, clones passing selection at 10 μg/ml of ampicillin were examined in order to identify a large-domain sequence that confers full resistance at this concentration (meaning no loss of colony formation; see Materials and Methods) but no resistance at concentrations not very much higher. The sequence chosen as the reference meets these conditions, conferring complete resistance at 10 μg/ml but none at 20 μg/ml (MIC = 20 μg/ml).

Relative to TEM-1, the reference sequence carries 33 substitutions scattered through the large domain, 29 of which are represented in the alignment (substitutions shown in boldface in Figure 6; see also Figure 7a). Substitution of key active-site residues in this sequence causes loss of function, indicating that the 10 μg/ml selection threshold is sufficient to eliminate sequences functioning by the uncharacterized mechanism encountered previously. Temperature sensitivity was assessed by repeating the ampicillin MIC measurements at 37 °C for strains producing no β-lactamase, the reference-sequence β-lactamase, or the wild-type TEM-1 β-lactamase. The resulting values (3.5, 4.0, and 4,200 μg/ml, respectively) give a reference-sequence activity of 0.01% relative to TEM-1 at 37 °C ((4.0–3.5)/(4200–3.5) = 10^{-4}). This is 30-fold lower than the 0.3% value measured at 25 °C ((20–5)/(5200–5) = 0.003), indicating that the reference-sequence enzyme undergoes substantial changes with increasing temperature in this range.

The hydropathy signature as a plausible fold-specific pattern

As is generally the case in experiments using the reverse approach (see Introduction), the fold adopted by functional sequences is restricted by the choice of experimental system. Here, because the function of the reference sequence traces back to the TEM-1 large-domain (with input from other large-domain sequences), we cannot expect other folds to be sampled in the randomization experiments. But, since many other folds might be comparably suitable scaffolds for this enzymatic function, how can we take this into account in our assessment of the overall prevalence of functional sequences?

Conceiving this prevalence as a fraction, the numerator would ideally be the number of sequences of large-domain length that provide a working β-lactamase (in the specified biological context) *via* any fold, and the denominator would

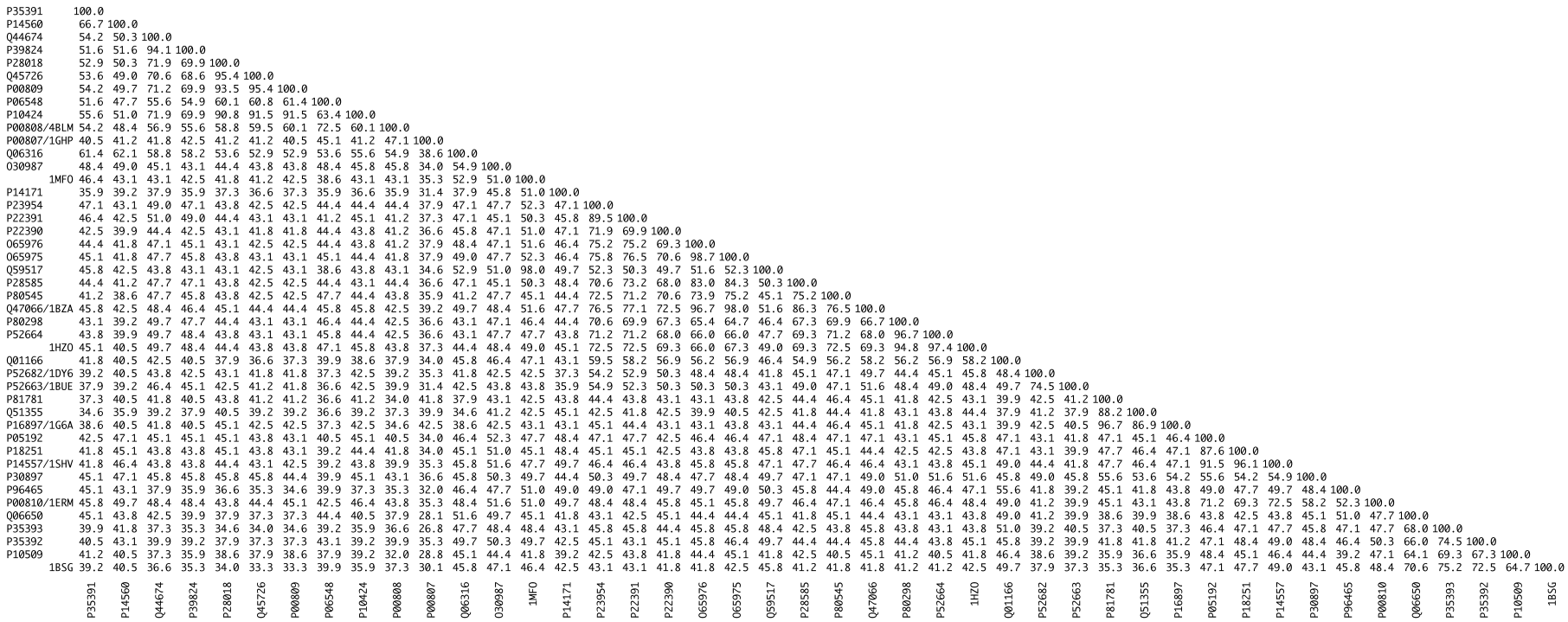

Figure 5. Sequence-identity matrix for the 44 aligned large-domain sequences. Residue identities (%) are based upon the full domain sequences (identified by SwissProt and/or PDB accession codes) as aligned in Figure 6.

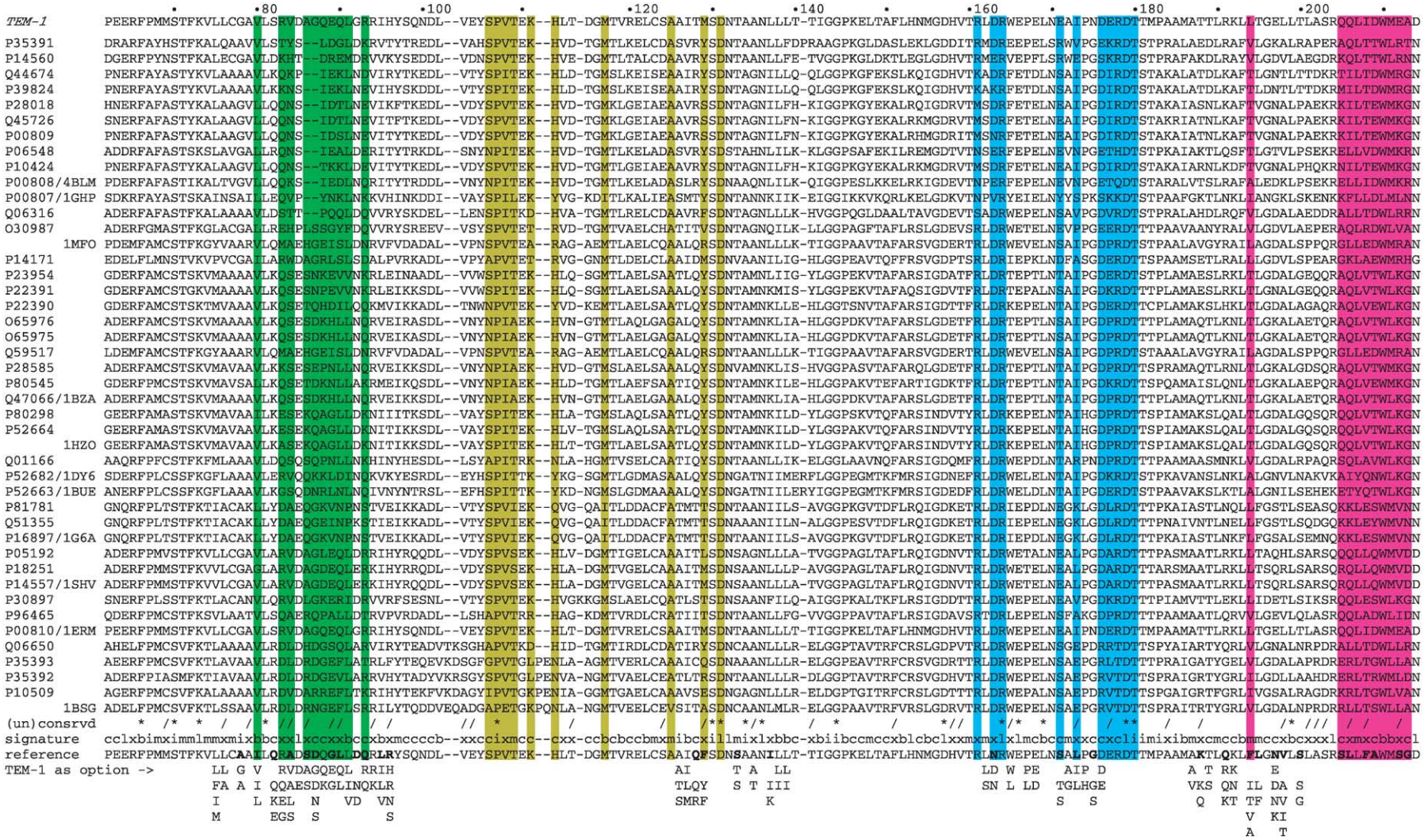

Figure 6. Alignment of 44 homologous large-domain sequences. Position numbering corresponds to the TEM-1 sequence, shown at the top and identified within the alignment by its SwissProt and PDB accession codes (P00810/1ERM). Shading indicates the four sets of ten positions chosen for randomization (coloured according to Figure 7b). Positions showing no variation are indicated below the alignment by asterisks (*). Those showing a high level of variation, meaning both a hydropathic constraint score of x (see Results and Discussion: subsection The hydropathy signature…) and six or more amino acid residues represented in the alignment, are indicated by slashes (/). Below the signature and reference sequences (explained in the text) are the allowed substitutions at the first four groups of positions subjected to limited substitution (see Results and Discussion, subsection Finding a reference sequence), the top row showing where the TEM-1 residue was included as an option.

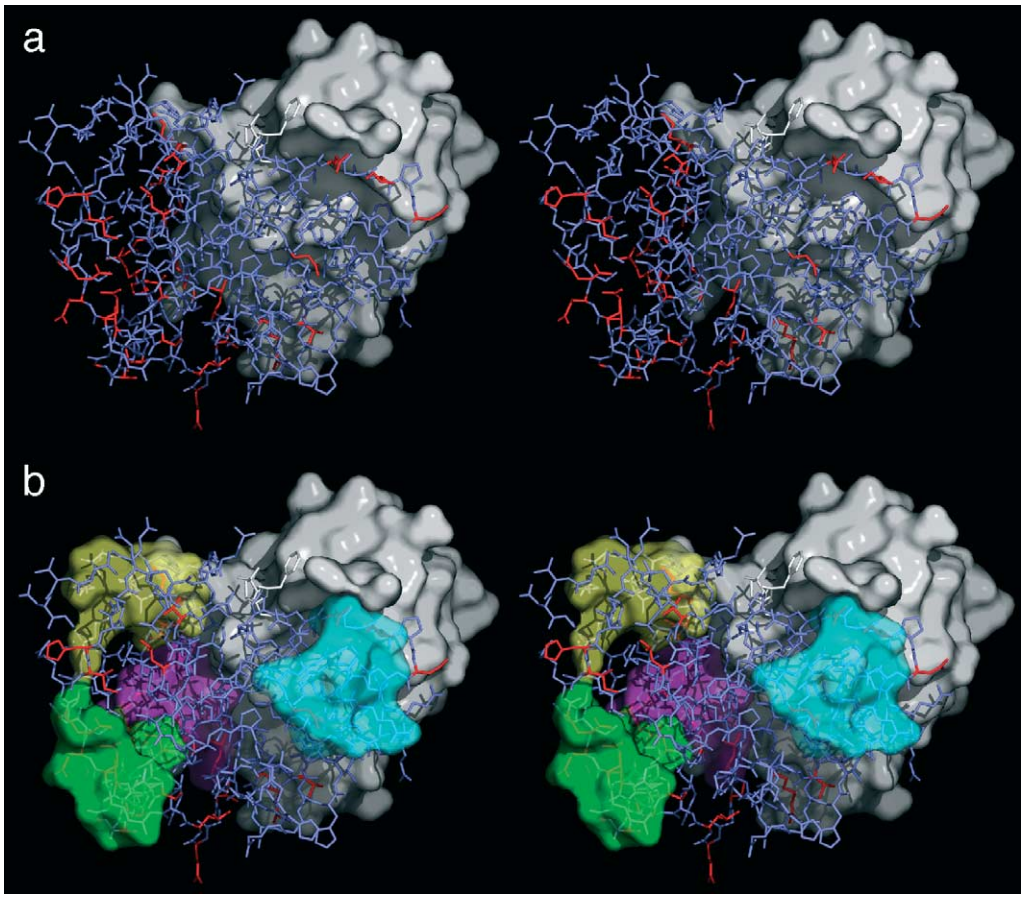

Figure 7. Location of reference-sequence substitutions and ten residue sets in the TEM-1 large domain (PDB 1FQG). Penicillin substrate (white) identifies the active-site pocket. a, Stereo image showing TEM-1 side-chains substituted in the reference sequence as red. b, Stereo image showing the four sets of ten residues chosen for randomization enclosed by transparent surfaces (see Results and Discussion, subsection Local side-chain randomization). Set 1 (green) includes positions 80, 83, 84, 86, 87, 88, 89, 90, 91, and 93; set 2 (gold) includes positions 106, 107, 108, 109, 111, 112, 117, 125, 129, 131; set 3 (cyan) includes positions 161, 163, 164, 171, 173, 176, 177, 178, 179, 180; set 4 (magenta) includes positions 194, 205, 206, 207, 208, 209, 210, 211, 212 and 213. Locations of reference-sequence substitutions are again indicated by red side-chains.

be the total number of possible sequences of this length. Realistically, though, the only numerator we can estimate by experiment is the number of sequences of large-domain length that provide a working β-lactamase *via* the large-domain fold. Still, we might hope to estimate the desired fraction by scaling the denominator appropriately. Instead of including all possible sequences of large-domain length, the scaled denominator should include only a fraction of these, that fraction being, to a first approximation, the inverse of the number of suitable folds.

This has direct implications for the design of the local randomization experiments, because the value of the denominator is effectively set in each experiment by specifying which amino acids are included as options at each randomized position. If all amino acids were included at all positions, we would be gathering data as though all of sequence space can be sampled meaningfully, whereas in reality we can sample meaningfully only the portion of space corresponding to the fold that has been fixed by the experimental system (the large-domain fold of Figure 4). Randomization should therefore be bounded in such a way as to restrict the sampling of sequence space to sequences that are inherently specific to that fold.

The fundamental role of the hydrophobic effect in the formation and stabilization of protein folds[12,20,21] may provide a means of doing this. For an amino acid sequence to encode a particular fold it is necessary, though clearly not sufficient,[8] that it favours burial of side-chains that will form the fold interior. This is achieved by means of an appropriate pattern of hydrophobic and hydrophilic residues along the primary sequence.[21] The causal connection between this pattern and the formation of folded structure, coupled with the geometrical connection between tertiary structure and the pattern of solvent exposure along the sequence, implies that folds should have highly specific

hydropathic requirements†. That is, apart from any consideration of physical interactions that depend upon the structures and precise orientations of individual side-chains, this more coarse interaction may be expected to severely limit the number of sequences that are compatible with a particular fold, different folds having distinctly different requirements.

The alignment shown in Figure 6 confirms this by providing clear evidence of conservation at the level of side-chain hydropathic character among sequences that show considerable variation at the level of side-chain identity. This can be seen by sorting the 20 amino acid side-chains into groups according to their hydropathic character and examining the alignment in terms of representation of these groups. Sixteen of the side-chains may be assigned to three groups as follows: hydrophobic group ≡ {F, L, I, M, V}, hydrophilic group ≡ {H, Q, N, K, D, E, R}, and intermediate group ≡ {G, S, T, Y}. These groupings are justified by chemical considerations (presence or absence of apolar surface, hydrogen bonding potential, or formal charge at physiologically relevant pH), by experimental measurements and theoretical estimates of free energies of transfer between water and apolar solvents,[22,23] and, to some extent, by the structure of the genetic code (all members of the hydrophobic group being specifiable by codons of the form NTN, and all members of the hydrophilic group being specifiable by codons of the form VRN; V indicating A, C, or G; R indicating A or G; N indicating any base).

Positions in the alignment may be placed into one of six hydropathic constraint categories according to representation of the above three groups: hydrophobic, hydrophilic, intermediate, not hydrophobic, not hydrophilic, or unconstrained (represented by the symbols `b`, `l`, `i`, `c`, `m`, and `x`, respectively). The four amino acids omitted from the above groups are best handled as special cases in this process. Two of these, alanine and tryptophan, are less hydrophobic than those of the hydrophobic group[22,23] but not uncommon at buried positions.[24] They are consequently best treated flexibly, according to the identities of other residues at the same position. Specifically, residues from the hydrophobic or intermediate groups, when present, will determine the constraint category. In cases where neither of those groups is represented, alanine and tryptophan will be interpreted as belonging to the intermediate group. The remaining two amino acids, proline and cysteine, introduce covalent backbone connections (intra-residue and inter-residue, respectively). Because this exceptional capacity is apt to be the determining factor in their placement, other side-chains should be given priority in assessing hydropathic constraints. When these principles are applied to the alignment (see Materials and Methods), hydropathic constraint scores by position are found to be as shown in Figure 6 (penultimate sequence).

As indicated above, physical considerations suggest that this sequence of constraint scores should be highly fold-specific, a unique signature of the large-domain fold. Two additional lines of reasoning support this. The first is based on the rarity of open reading frames encoding sequences consistent with this signature. This may be estimated from the constraint scores by taking insertions and deletions (indels) into account. Because these mutations expand or contract the backbone, they are expected to be highly disruptive at most locations. This is confirmed by the alignment shown in Figure 6, and by other studies of natural variation in coding sequences.[25] The natural large-domain variants show indels at five points that cluster on the exterior of the folded structure, on the face opposite the interface with the smaller domain. All occur at highly exposed locations either in turns or near the ends of short, peripheral helices. Consistent with this, the optional positions are filled predominantly by hydrophilic residues or by proline residues (Figure 6). In view of the total number of DNA base changes represented throughout the alignment (of the order of 10^3), the paucity of indels along with their common structural features is clearly indicative of functional constraints. That one of the few represented indels appears to have two independent origins (after position 140) further suggests that the represented set is nearly complete.

Assuming that the represented indels may be tolerated in any combination, we may estimate the proportion of open reading frames carrying the large-domain signature to be about 10^{-33} (see Materials and Methods). If this is smaller than the inverse of the estimated total number of possible folds, that would indicate that the signature is sufficiently restrictive to be fold-specific. Despite considerable uncertainty as to the total number of possible folds, there is an emerging consensus[26–29] that fundamental constraints on protein structure limit the figure to something very much smaller than 10^{33}, which implies that the signature is amply restrictive to be fold-specific.

Secondly, as an empirical test of fold specificity, we can determine whether any known proteins unrelated to β-lactamases come close to fitting the large-domain signature. To do this, the signature was divided into three sections, each spanning 51 positions. A pattern search was then used to examine the human, fly, worm, and yeast proteomes‡ for sequences fitting any of these

† "Fold" is here taken in the tight sense exemplified by the large-domain fold (Figure 4). Although fold similarities much less tight than those of Figure 4 may indicate homology, position-by-position properties and constraints vary considerably as similarities become more loose.[44] Still, hydropathic constraints remain evident so long as there is tight structural similarity over a sizeable portion of structure.[45]

‡ http://www.ensembl.org/Homo_sapiens; http://www.flybase.org; http://www.wormbase.org; http://www.yeastgenome.org

Table 1. Characteristics of the ten residue sets

	Substitutions in reference[a]	Conserved positions[b]	Diverse positions[b]	Buried positions[c]	Exposed positions[d]
Set 1	7	0	5	3	3
Set 2	1	2	1	6	1
Set 3	3	3	2	4	1
Set 4	7	0	3	5	4
Set average	4.5	1.3	2.8	4.5	2.3
Expected[e]	2.2	1.1	2.1	5.2	2.5

[a] Substitutions carried by the reference sequence (relative to TEM-1) within the specified set.
[b] See Figure 6.
[c] Side-chains in the TEM-1 structure having less than 20% maximal solvent exposure, as calculated by GETAREA 1.1 (hhtp://www.scsb.utmb.edu/getarea/).
[d] Side-chains having greater than 50% maximal solvent exposure (see footnote **c**).
[e] Expected values for ten randomly chosen positions in the large domain.

signature sections. Since proteins known to have the large-domain fold or a clearly related fold are all of prokaryotic origin, they cannot appear as matches. None of the proteome sets, in fact, shows matches to any of the three sections, indicating a high degree of signature specificity in this empirical sense.

Local side-chain randomization

Four sets of residue positions in the reference sequence (coloured in Figure 6) were chosen for separate randomization experiments. Each set comprises ten residues in close proximity in the native large-domain fold (Figure 7b). Variants from each of these experiments that enable colonies to grow in the presence of 10 μg/ml of ampicillin show themselves to have adequately fold-favouring side-chain interactions within the randomized regions. In principle, the whole large domain could be examined with 15 such experiments, each covering about ten residues. In practice though, the positions involved in each experiment must be sufficiently close in sequence that their codons can be spanned with a pair of oligonucleotide primers (see Materials and Methods). The four chosen sets meet this condition and together cover a significant fraction (26%) of the fold.

Comparison of these sets to the whole domain shows them to be reasonably representative in terms of the average frequency of various position-specific attributes (Table 1). However, they are clearly skewed toward greater inclusion of substituted positions in the reference sequence (first column). This has been arranged as a means of erring on the side of caution for the following reason. Since the reference sequence has been produced in such a way that it carries nearly as much structural disruption as it can bear under the specified test conditions, and this disruption was caused by departure from the TEM-1 sequence at 22% of the large-domain positions, we expect it to be more sensitive to further changes within the 78% that match TEM-1 than to alternative changes within the 22% that differ. In other words, changing what has already been changed is less apt to cause further disruption than changing what has been retained. In particular, pivotal residues (see Experimental Approach) are distributed among the unaltered 78% in a manner that cannot be predicted reliably. The best way to guard against accidental over-representation of such residues among randomized sets, thereby guarding against exaggeration of the sequence constraints, is therefore to include a disproportionate number of positions from the altered 22%.

In designing the randomizing primers (see Materials and Methods), the large-domain signature was used to restrict the explored sequences to those conforming to the hydropathic requirements of this fold. As discussed above, the purpose of this is to limit the sequence possibilities in a manner that is consistent with the one-fold structural limitation. Randomization was performed first at set 4, with one of the resulting variants chosen as the reference (see Results and Discussion, subsection Finding a reference sequence). This reference sequence was then used as the starting point for the subsequent experiments. In each experiment, the prevalence of working sequences among valid test sequences (the pass rate) is determined from colony counts and the measured frequency of invalid constructs (Table 2). Ampicillin-resistant colonies were found in two of the four experiments (Table 2, column 10), enabling clear quantification of pass rates. Upper-bound estimates of pass rates are attainable from the other two experiments, and in one of these (set 3) isolation of a few working sequences in the initial library shows this estimate to be close to the actual figure.

Several of the randomized genes found to confer ampicillin resistance were sequenced in order to look for any clear patterns (Figure 8). One interesting observation is that side-chain conservation seen at this low functional threshold shows some departure from conservation among the natural homologues. The threonine residue at position 180, for example, is invariant among the homologues (Figure 6) but replaceable in the reference sequence. Conversely, the homologues have leucine as often as methionine at position 211, but methionine appears to be preferred decisively among the functional randomized sequences. Also, although

Table 2. Calculation of pass rates from local side-chain randomization experiments

Set	Column										
	1	2	3	4	5	6	7	8	9	10	11
	Initial library size[a] (k)	Colonies on chlor controls[b]	$Chlor^R$ cells per test plate[c] (k)	$Chlor^R$ cells tested[d] (k)	Gross test size[e] (k)	Junction pass rate[f]	Sequence pass rate[g]	% Signature-consistent[h]	Net test size[i] (k)	Amp^R colonies[j]	Amp^R pass rate[k](%)
1	330	138, 166	61	370	330	19/30	7/10	85.3	125	41	0.03
2	240	130, 150	56	340	240	19/30	5/10	70.4	54	0	<0.002
3	600	154, 158	62	370	370	20/30	5/11	90.9	102	0	~ 0.001
4	540	122, 111	47	470	470	14/30	3/10	90.9	60	18	0.03

See Materials and Methods for a full description of the calculation.

[a] Based upon colony counts on chloramphenicol plates (20 μg/ml) following the initial post-mutagenesis transformation.

[b] Counts for two chloramphenicol plates (7 μg/ml), each spread with 20 μl of a 10^{-6} dilution of the saturated test cultures.

[c] Chloramphenicol-resistant cells spread onto each ampicillin test plate, calculated by multiplying the ratio of dilutions (200) by the sum of the counts in column 2 (control spreads being half the volume of the test spreads).

[d] From column 3 and the number of ampicillin test plates in each experiment (six for sets 1–3; ten for set 4).

[e] The lower of the numbers in columns 1 and 4.

[f] Results of restriction analysis performed on plasmids prepared from 30 control clones (column 2) from each experiment.

[g] Results of DNA sequence analysis performed on plasmids that passed the junction test (column 6).

[h] Calculated from the number of NNK codons (5, 2, 3, and 3) and VRW codons (0, 1, 1, and 0) in the respective experiments.

[i] Calculated from gross test sizes (column 5) by multiplying by the three fractions in columns 6–8.

[j] Total counts on six test plates for sets 1–3, or ten for set 4.

[k] From ratio of numbers in columns 10 and 9, but using a minimum count of one colony. Although no colonies appeared on the ampicillin test plates for sets 2 or 3, thorough screening of the initial libraries (column 1) revealed a few Amp^R clones in the set 3 library.

	set 1	set 2	set 3	set 4
	•89 •93		•161 •163 •180	•205 •209 •211 •212
TEM-1	VRVAGQEQLR	SPVTKHMAMD	RDREIDERDT	LQQLIDWMEA
reference	I**R**ASD**Q**GL**L**Q	**SPVTKHMA**F**D**	**R**N**R**SL**DERDT**	FSL**L**FA**WM**SG
signature	bxxxccxxbc	cixmccbmxl	xxlcxcxcli	mcxmxcbbxc
	VRVV**G**EWS**L**D		**RDRE**TAVE**DT**	VH**QL**LG**WMCA**
	VRCR**G**NLK**LT**		NA**R**DI**D**L**RDT**	**L**KI**L**MG**WM**CD
	VG**V**N**G**TMA**L**H		**RDR**QAD**T**R**D**S	**L**ST**L**FA**WM**CG
	VASE**G**KLT**L**S			**L**DM**LL**A**WM**IG
	IY**V**MQKLD**L**A			IDV**LL**A**WM**TR
	I**R**IRREMGIQ			FSL**L**FA**WM**SG

Figure 8. Functional residue combinations identified from four separate randomization experiments. Wild-type TEM-1 residues, reference-sequence residues, and signature scores are listed for each set in ascending order according to position (see Figure 6 or Figure 7 for unlabeled position numbers). Functional combinations found by sequencing complete genes of passing clones are listed below the signature scores, with matches to the TEM-1 sequence shown in boldface. For sets 1 and 4, these clones were among those counted in the assessment of pass rates (Table 2). The three functional combinations shown for set 3 were isolated as described (Materials and Methods) after the initial selective plating produced no colonies. No functional variants could be isolated following randomization of set 2.

there is a clear tendency toward preservation of, or reversion to, TEM-1 residues among the functional variants (and this cannot be attributed to template bias because randomized regions are introduced as insertions; see Materials and Methods), there are intriguing deviations from this. For example, position 212, which carries a glutamate residue in the TEM-1 sequence context, seems to be suited to the cysteine side-chain in many of the contexts explored here†.

Some of the features of the functional sequences are explicable with reference to the TEM-1 structure. Position 89, occupied by a glutamate residue in TEM-1, shows a distinct preference for bulky, apolar side-chains in the randomized variants (Figure 8). In TEM-1, Glu89 forms a salt-bridge with Arg93. A number of the natural homologues have similar salt-bridges between glutamate residues at position 89 and either lysine or arginine residues at position 93. The randomized variants, on the other hand, seem to accommodate wider variation at position 93 by placing relatively large and hydrophobic side-chains at position 89. A similar situation seems to occur within set 4, where Gln205 forms a hydrogen bond with Asp209 in the TEM-1 structure. The mutants appear to accommodate a variety of side-chains at position 205 by truncating the aspartate side-chain to glycine or alanine. The Arg161–Asp163 salt-bridge in set 3, while not fully conserved, is much more dominant in the alignment than the previous examples. Although both positions are scored *x* in the signature, more restrictive randomization was used to favour arginine at position 161 (see Materials and Methods). Despite the lopsided likelihood of receiving the respective residues upon randomization (25% chance of Arg161 *versus* 3% chance of Asp163) they appear together or not at all in the isolated variants (Figure 8). These examples all suggest that charged side-chains, while clearly capable of improving the ability of a fold to deliver function, tend to offer this benefit at the cost of rather particular contextual requirements.

Together with the pass rates, the prevalence of TEM-1 residues among functional variants appears to confirm the expected relationship between a set's degree of substitution in the reference sequence and its tolerance of randomization. Sets 1 and 4, both 70% modified (relative to TEM-1) in the reference sequence, show functional variants averaging 68% and 62% modification (Figure 8). The acceptability of these high levels of modification appears to correlate with relatively high pass rates (Table 2), even though there is clear evidence of selective constraints at modified positions. Set 3, only 30% modified in the reference sequence, shows significantly lower modification among functional variants (40%) and a significantly lower pass rate. It seems likely, then, that the inability to isolate functional variants following randomization at set 2 is related to the low degree of modification (10%) of these positions in the reference sequence.

Implications

The exponential relationship between possible sequence combinations and chain length makes exhaustive experimental searching of sequence space impossible for anything but small peptides. Simplifying assumptions will therefore always be essential for treatments of the spaces corresponding to proteins of biological significance. Yet, given the importance of these concepts to our understanding of such basic things as protein folding, stability, and evolution, the difficulty of achieving anything like certainty should not deter us from exploring the validity of such assumptions. Since they need not be provable to be testable (i.e. disprovable), we can reasonably hope for convergence upon correct ideas through a succession of testable hypotheses.

For the purposes of the present study, it seems reasonable to assume that the pass rates of Table 2, when averaged, provide an upper-bound estimate of the true mean pass rate (i.e. the mean that would result from applying the same method to sets of ten residues that cover the entire domain). Several aspects of the analysis justify this. First, one of the four pass rates is itself an upper-bound estimate, no

† The reference sequence and all randomized variants retain the pair of cysteine residues at positions 77 and 123 (Figure 6) that form a disulfide bond in the TEM-1 structure. Position 212 is not in the vicinity of the disulfide in this structure.

functional variant having been found for set 2. Second, as described above, randomized sets were made to include more than a representative number of substituted positions in the reference sequence as a way of avoiding exaggeration of constraints. And third, the pair of cysteine residues that forms a disulfide in the TEM-1 large-domain has been retained, potentially enabling this fold-favouring bond to form in randomized variants. The per-position geometric mean calculated from the four pass rates is 0.38:

$$[(3\times10^{-4})\times(2\times10^{-5})\times(10^{-5})\times(3\times10^{-4})]^{1/40} = 0.38$$

By the above assumption, this should be interpreted as an upper-bound estimate of the mean likelihood that a side-chain that complies with the large-domain signature will form adequate interactions with neighbouring signature-compliant side-chains.

How is the overall prevalence of adequacy among large-domain sequences fitting the signature related to this per-position likelihood? In other words, what pass rate should we infer for an ideal experiment in which the whole large domain is simultaneously randomized within the constraints of the signature? In answering this it will be helpful to consider some fundamental aspects of the relationship between amino acid sequence and tertiary structure.

Protein folding is a cooperative process[21] in which a large number of weakly fold-favouring interactions combine to cause a concerted transition to the folded state. Although the main chain is involved in many of these interactions, it is the side-chains that must account for the causal connection between sequence and structure. In a folded protein, each side-chain is surrounded, fully or partly, by a particular set of protein atoms with which it must interact directly. Although there is no theoretical distance at which direct pair-wise interactions cease, interactions with close neighbours will dominate for a number of physical reasons (e.g. the inverse-square nature of coulombic forces, charge screening, and limits to lengths of hydrogen bonds). We may therefore think of the overall problem of fold stabilization as consisting of a collection of coupled local problems. Each of these local problems is solved by specifying side-chains that adequately favour the native conformation locally. Coupling results from the fact that most of the local problems cannot be solved separately. The reason for this is simply that few associations between residues distant in sequence (i.e. tertiary contacts) can be made so favourable as to be formed decisively even when the rest of the chain is unfolded. But the local problems become progressively more tractable as the number of accessible non-native states is narrowed progressively (the folding funnel principle[21,30]). Consequently, the whole collection of local problems tends to be solved jointly (over domain-sized regions) or not at all.

So, for a randomized variant from the above ideal experiment to pass selection, side-chain specifications would have to provide such a joint solution of local problems throughout the large domain. Since the four randomization experiments provide an upper-bound estimate of the likelihood of solving the local problem for a ten-residue set, the likelihood of the joint solution may be estimated by applying the above per-position mean (0.38) across the domain. The resulting figure, 10^{-64} ($0.38^{153}=10^{-64}$), is thus an upper-bound estimate of the prevalence of functional sequences among the whole set of signature-compliant large-domain sequences.

How does this compare to estimates from earlier work on other proteins that used the reverse approach? We can adjust the figure to obtain a rough estimate of the prevalence of functional large-domain sequences among all sequences of this size (signature-compliant or not). To do this, we multiply by 10^{-33}, the estimated proportion of all open reading frames that encode the large-domain signature (see above, and see Materials and Methods), resulting in a figure of 10^{-97}. Reidhaar-Olsen and Sauer[7] estimated the proportion of 92 residue sequences that form a functional λ-repressor fold to be 10^{-63}. When scaled according to chain length, this gives 10^{-105} as the corresponding proportion for a 153 residue fold ($10^{-63(153/92)}=10^{-105}$). As they indicated,[7] their assumption of context independence leads to overestimation of the working proportion. Their high selection threshold (5 to 10% of wild-type activity) has the opposite influence, the net effect being quite good agreement with the present result as well as with earlier calculations[6] based on natural variation in cytochrome *c*.

As discussed in Introduction, the method applied in the study of chorismate mutase by Taylor and co-workers[9] should provide a more accurate estimate than the earlier λ-repressor study. Their search for functional chorismate mutases was restricted to sequences matching the hydropathic pattern of a natural version of the enzyme. So, bearing in mind the difference between a single-sequence pattern and a multi-sequence signature, their estimated functional prevalence should be compared to the estimated prevalence among signature-compliant sequences in the present study. Scaling their figure gives 10^{-40} for a 153 residue sequence ($10^{-24(153/93)}=10^{-40}$). This is significantly larger in logarithmic terms than the above estimate for the large domain (10^{-64}). However, in view of the difference in fold complexity (Figure 1) and the fact that pattern-based randomization is more restrictive than a signature-based randomization, there is no reason to think the two estimates are inconsistent. It seems, rather, that a number of studies using the reverse approach lead to a consistent picture in which sequences with function clearly akin to that of natural proteins are extremely rare, the exact

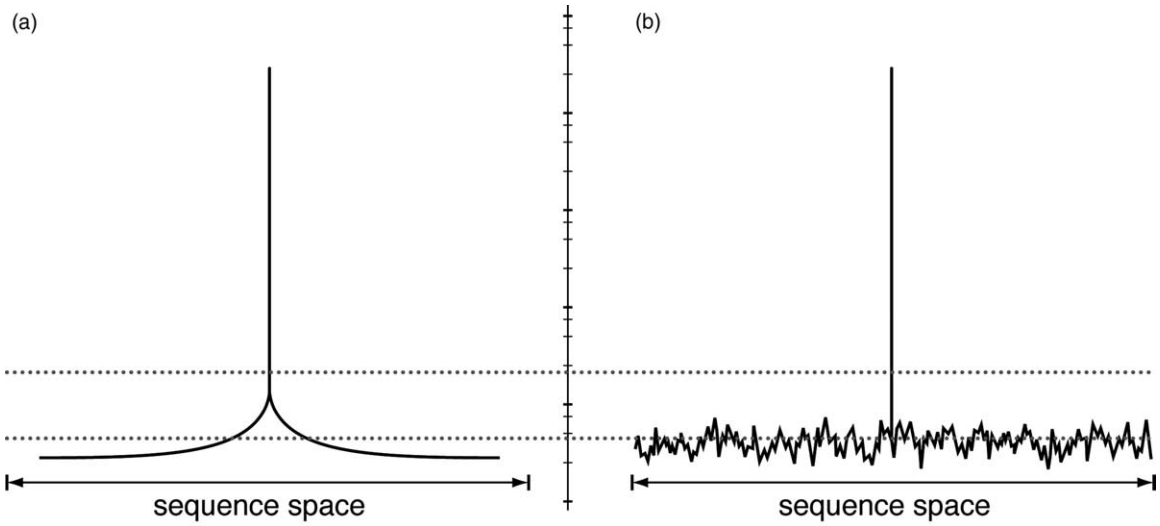

Figure 9. Alternative models of how function might map onto sequence space. A quantifiable function, performed with a high level of efficiency by a natural protein, is represented in the vertical dimension, the logarithmic scale indicating a wide range of measurable activity. For the purpose of this comparative illustration, sequence possibilities are imagined to be represented within the horizontal scales such that neighbouring sequences occupy neighbouring positions on these scales. Dotted lines represent two selection thresholds, function being much rarer at the higher threshold. (a) Global-ascent model, with optimal sequence in the middle. (b) Local-ascent model, with optimal sequence in the middle.

degree of rarity depending upon the complexity of the fold.

How might this picture be reconciled with the much higher prevalence of function often reported in studies using the forward approach? Figure 9 illustrates two possible ways for functional sequences to appear relatively common when a very low functional threshold is used. Figure 9(a) represents a global-ascent model of the function landscape, meaning that incremental improvement of an arbitrary starting sequence will lead to a globally optimal final sequence with reasonably high probability. In this case, sequences exhibiting function at any level are properly regarded as suboptimal versions of the optimal archetype. Consequently, if we want to know how common sequences of this functional type are (regardless of optimality), we should set the functional threshold as low as possible. The higher of the two thresholds shown in Figure 9(a) would therefore lead to a considerable underestimate. However, if the real landscape is more like the local-ascent model depicted in Figure 9(b), where incremental improvement leads to an archetypal sequence for only a relatively tiny set of local starting sequences, then the lower threshold would lead to a considerable overestimate. In essence, activity might be a reliable marker of archetype-like mechanism down to some minimum level, but not below.

Considering that the functional mechanisms of natural proteins are intrinsically dependent upon well-defined tertiary structures†, a reasonable hypothesis is that activity ceases to be a reliable marker of native-like mechanism at the point where it is low enough not to require something akin to native-like tertiary structure. The present study takes advantage of two functional sequences, one that employs the known enzymatic mechanism and one that does not, in order to set the functional threshold at a level that seems to require a working active site. Since formation of the active site requires tertiary structure of some sort, by merely requiring a working active site, we ensure that we are focusing on the relevant sort of structure: i.e. what is needed for a crudely functional enzyme fold. Modes of catalysis that do not require this sort of structure, however real and interesting they may be in some respects, do not explain how this sort of structure appears as new folds emerge.

Because forward-approach studies showing function to be much more prevalent than indicated here do not report tertiary structure,[3–5] the possibility that the reported functions might not require such structure must be considered. The fact that peptides too small to fold may bind ligands,[31] and even show some catalytic activity,[17] shows that these functions do not necessarily imply folded structure. Similarly, larger proteins may avoid proteolysis *in vivo*,[32,33] exhibit cooperative thermal denaturation,[34] and even possess catalytic activity[32] without having native-like tertiary structure. Indeed, considering the difficulty encountered in concerted efforts to design native-like structure into very simple folds,[35] it would be surprising if such structure were prevalent in random sequence libraries. In light of all the available evidence, then, Figure 9(b) seems to

† This is not to say that the functional structure must always form independently of substrate/ligand binding,[46] but merely that functional mechanisms of natural proteins are invariably explicable in terms of defined structures.

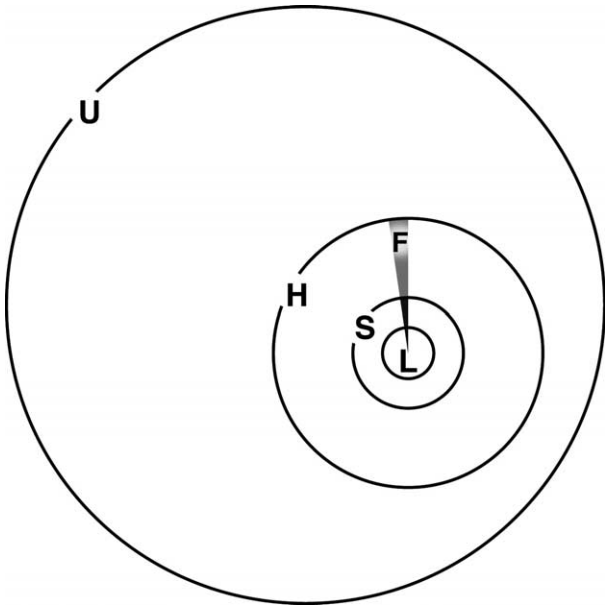

Figure 10. Relationships between key sets and subsets of sequences. See the text for a full explanation.

offer the more plausible way to reconcile the findings of forward-approach studies with the findings of reverse-approach studies.

If we provisionally take this to be the case, we may use the estimates of the reverse approach to obtain at least a tentative figure for the overall prevalence of sequences of a given length that perform a particular function by means of proper tertiary folds. For the present purposes, we take the length and function of interest to be those of the large domain. Figure 10 illustrates the relationship between the relevant sets of sequences in terms of a Venn diagram. The full set of possible sequences of large-domain length is represented by points within the outer circle (U signifying unconstrained sequences). Some fraction of these sequences, represented within the H circle, meet the hydropathic requirements for specifying one of the many possible tertiary folds of this size. Possibly many of these folds are capable of complementing the smaller TEM-1 domain to form a properly functioning β-lactamase. Points within the S circle correspond to sequences meeting the hydropathic requirements for forming these suitable folds. Sequences complying with the hydropathic requirements for one such fold, the large-domain fold, are represented within the L circle. Sequences within the shaded sector, F, not only carry the hydropathic pattern of a fold but also provide the fold-favouring interactions needed to stabilize that fold (folding sequences). The desired prevalence is represented by the size of the intersection of F with S (shown as the dark portion of F) relative to the size of the whole set, U.

The proportion of folds suited to the specified function (corresponding to the proportion of points within H that are also within S) can be estimated roughly by considering the question of fold suitability more generally. The historical likelihood of existing folds being suited to new functions may be inferred from the number of distinct fold types in nature, where type here refers to a set of folds of sufficient similarity that they may plausibly be attributed to recruitment or divergence from a common progenitor. Since recruitment of an existing fold type to serve a new function is easier than generation of a new type, we expect recruitment to occur whenever it is feasible. Consequently, if the total number of fold types in use is of the order of 10^4 (see Coulson & Moult[36]) with something like 10^3 employed in individual species,[37] this gives us an idea of the number of fold types required to cover all biological functions. A reasonable estimate of the average proportion suited to a particular task is 0.1%. This would enable a set of 4000 fold types to provide 98% coverage of functions ($1–0.999^{4000} = 0.98$) and a set of 8000 to provide 99.97% coverage.

Based upon the estimated proportion of set L that is within sector F (10^{-64}, as above) and in view of the scaled figure from the chorismate mutase study (10^{-40}), we may estimate that sector F subtends something in the range of one part in 10^{64} to one part in 10^{40} of circles H and S. What proportion of all sequences (set U) fall within set H? Lau & Dill have carried out a theoretical analysis of foldability based upon hydropathic constraints alone.[38] Their least stringent folding criterion gives a value of 10^{-10} for this proportion, which would mean that of all sequences in U, something like one in 10^{74} to one in 10^{50} form folded structures (i.e. fall within F). So, if set S is about one-thousandth the size of set H (as above), then the proportion of all sequences of large-domain length that perform the specified function by means of any tertiary fold (i.e. fall within the dark portion of F) is estimated to be in the range of one in 10^{77} to one in 10^{53}.

At first glance, it seems implausible that natural sequences could diverge through a space where function is represented so sparsely. How, for example, can we account for the substantial diversity among the large-domain homologues (Figure 5) if randomly altered sequences have such slim prospects of retaining function? The answer follows from the fact that functional sequences are not distributed uniformly through sequence space. A random change to a functional sequence actually has a good chance of leaving function undisturbed if very few positions are affected. As estimated above, the likelihood of a signature-compliant substitution in the large-domain reference sequence producing a comparably functional variant is about 38%. Since 70% of the ~1000 possible non-synonymous base changes to the reference coding sequence produce signature-compliant substitutions, about one in four random single-residue changes are functionally neutral. The proportion would be somewhat lower under conditions requiring a higher level of function (such as those under which neutral drift normally occurs) but not so low as to preclude progressive

sequence divergence by gradual accumulation of point mutations.

However, it is not obvious that fold diversity is as easily explained as sequence diversity, if functionally folded sequences are as rare as this analysis indicates. A commonly accepted view is that new folds are pieced together from small parts of existing folds.[32,33,39,40] But to the extent that a new fold is really new, its formation must require the joint solution of at least a considerable number of new local stabilization problems of the kind described above. How likely is it that sequences that carry the hydropathy signatures of other folds and provide joint solutions to the stabilization problems for those folds may be pieced together in such a way that they satisfy a new set of constraints, equally demanding but substantially different? The analysis provided here, bearing in mind the uncertainties, calls for careful examination of such piecing scenarios. The need for caution is underscored by a recent study of the structural and functional consequences of piecing together parts from homologous versions of the same fold.[41] Because even close homologues employ substantially different solutions to their local stabilization problems,[8] chimeras made by homologous recombination suffer considerable disruption unless the points of crossover minimize intermixing of these local solutions.[41] So, if re-creating a fold by ordered assembly of sections of sequences that already adopt that fold is not a simple matter, generating new folds from parts of old ones may be much less feasible than has been supposed.

Materials and Methods

Large-domain sequence alignment

The FASTA algorithm was used with the blosum50 scoring matrix to search the SwissProt database for sequences at least 50% identical with the large-domain sequence of any of the 11 structural representatives. Sequence identity was judged over the entire length of the domain. The resulting set of sequences contains several single-position variants of the SHV-1 (SwissProt P14557; PDB 1SHV) and PSE-4 (SwissProt P16897; PDB 1G6A) large domains, which were removed to minimize redundancy. The SwissProt entry for the Toho-2 enzyme (O69395) was also removed on the grounds that it deviates radically from the others over a region of about 35 positions,[42] including one that is otherwise invariant. Examination of the reported gene sequence for Toho-2 shows that a sequence very similar to that of Toho-1 (SwissProt Q47066; PDB 1BZA) can be reconstructed with a few base insertions in this region. Given the apparent improbability of a series of point deletions (affecting a third or more of the enzyme) passing natural selection, and since the possibility of errors in sequencing or point deletions occurring during subcloning cannot be wholly excluded, it is prudent to remove the Toho-2 sequence. The final set of 44 sequences was aligned using the CLUSTAL W algorithm initially, with structural comparisons used to make minor adjustments.

Obtaining the hydropathy signature

The procedure is outlined in the text. In order to minimize the possibility of sequence errors affecting the signature, a hydropathic group is counted as being represented at a position in the alignment if it is represented in any of the 11 structures or in at least two sequences that lack structures. Where the two extreme groups (hydrophobic and hydrophilic) are represented on this basis without representation of the intermediate group, a score of `x` (no hydropathic constraint) is assigned. Position 107, occupied exclusively by proline in the alignment (Figure 6), is the only position where this procedure does not produce a definite score (owing to the special treatment of proline, described in the text). Because this position is included in one of the randomized sets (set 2), a score needs to be assigned. Pro107 marks a loop-to-helix transition in the large-domain fold. The likely role of the proline side-chain in preventing extension of the helix suggests that hydropathic constraints may be of secondary importance here. But because proline has intermediate hydropathic character,[23] and often aligns with residues of intermediate hydropathic character,[43] and because position 107 shows an intermediate degree of solvent exposure (25%), this position is scored as `i`. This interpretation provides maximal representation of proline in the variants produced by randomization of set 2.

Estimating the proportion of sequences carrying the signature

All size comparisons between sets of sequences in this work assume a codon basis, meaning that the absolute sizes may be interpreted as the total number of distinguishable coding sequences. Fifty of the 61 sense codons encode residues with unambiguous hydropathic character, according to the three groups defined in the text (see Results and Discussion, subsection The hydropathy signature…). Although the remaining 11 (encoding Ala, Trp, Pro, and Cys) are less clear-cut, we can obtain a reasonable estimate of the desired proportion by dividing these among the hydrophobic and intermediate groups, reflecting their actual position on the scale.[22,23] In this way, we allocate 21, 18, and 22 codons to the hydrophobic, hydrophilic, and intermediate groups, respectively. This gives the following numbers of codons complying with each of the six hydropathy scores: 21 for `b`; 18 for `l`; 22 for `i`; 40 for `c`; 43 for `m`; 61 for `x`. So, the proportion of open reading frames encoding proteins that are consistent with a specified score sequence is calculated as the product $(21/61)^{b}(18/61)^{l}(22/61)^{i}(40/61)^{c}(43/61)^{m}$, where exponents are the number of occurrences of the respective scores.

The signature corresponding to a tightly defined fold will be more complex than a simple score sequence if multiple indel variations are consistent with that fold, as is the case for the large domain. To account for this, the alignment shown in Figure 6 may be divided into six nonoverlapping blocks, the first consisting of positions 62–85, and subsequent ones starting at successive indel locations. The likelihood of an open reading frame comporting with the full signature may then be calculated from separate calculations on each block that treat indels as optional prefixes. The resulting figure is one in 10^{33} (for details of the calculation, see the Supplementary Material).

Plasmids and strains

The starting plasmid for this work was derived from pUC18 by inserting the *cat* gene (conferring resistance to chloramphenicol) at the HindIII site and replacing the XmnI to AlwNI fragment (1 kb) with the corresponding fragment from plasmid pBR322. This replacement corrects two missense mutations in the β-lactamase gene (*bla*) carried by pUC-type plasmids, restoring the encoded sequence to that of the wild-type TEM-1 enzyme (SwissProt P00810). *Escherichia coli* strain TOP10 (Invitrogen) was used in all experiments. Oligonucleotides were synthesized and PAGE-purified by SIGMA-Genosys.

Quantitative ampicillin selection protocol

Several precautions were taken in the measurement of MIC values and in applying ampicillin selection at precise threshold levels. To avoid irregularities that may occur when cells are spread on ampicillin-containing medium at high densities, very dilute cultures were spread. Also, to prevent any accumulated selective history of cell lines, all cells used in critical selection work were encountering ampicillin for the first time. Where necessary, TOP10 was re-transformed with prepared plasmid to obtain a strain lacking a history of ampicillin exposure.

In preparing plates for ampicillin selection, molten LB-agar medium was equilibrated at 54 °C prior to addition of freshly prepared ampicillin solution. Plates were poured the day before use. On the day of the experiment, cultures grown at 37 °C in 2× TY medium containing chloramphenicol (20 μg/ml; no ampicillin) were diluted 5000-fold (or 10^6-fold for full-resistance check) in LB medium, and immediately spread on the selective plates (40 μl per 90 mm plate or 20 μl for full-resistance check). In addition to ampicillin, these plates contained chloramphenicol at a concentration of 7 μg/ml. Each test culture was also spread (20 μl of 10^{-6} dilution) on ampicillin-free plates containing 7 μg/ml of chloramphenicol. Wrapped plates were incubated at 25 °C for 42 hours (or 20 hours where 37 °C incubation is indicated).

Measurements of MIC were performed by serial plating with ampicillin increasing in increments of 0.5 μg/ml up to 7 μg/ml, 1.0 μg/ml up to 12 μg/ml, 2.0 μg/ml up to 24 μg/ml, and in 200 μg/ml increments for measurements of wild-type TEM-1 activity. MIC is taken to be the lowest level showing no visible growth at the end of the incubation period. To check the reference-sequence strain for full resistance to 10 μg/ml of ampicillin (at 25 °C), colony counts were compared on plates having or lacking ampicillin at this level (both having 7 μg/ml of chloramphenicol). On two plates without ampicillin, the counts were 137 and 141. On two plates with ampicillin the counts were 136 and 139, indicating no detectable loss of colony formation.

Insertion mutagenesis

All mutagenesis steps in this work, for producing the reference sequence or randomizing a set of positions within it, involve the same basic steps. First, PCR using outwardly directed primers is used to delete the entire region spanning the codons to be substituted, leaving a unique restriction site in place of the flanking codons. Then, following cleavage at that site, mixed-base oligonucleotides (outwardly directed) are used in a second PCR to restore the full-length open reading frame. This makes it possible to select for ampicillin resistance following mutagenesis without any background from unmodified template DNA, and it prevents bias from the initial template at the points of substitution.

Production of the reference large-domain sequence

Three rounds of insertion mutagenesis covering 39 amino acid positions were performed in succession (i.e. without transforming cells at intermediate stages). The initial template was a plasmid in which all three regions are deleted from the TEM-1 *bla* gene. The amino acid sets shown at the bottom of Figure 6 (first three groups) were represented in the oligonucleotide mixtures used in the successive insertion steps. Representation of both Ala and Leu at position 76 required combining separately synthesized primers. The final ligated product was used to transform *E. coli* strain TOP10 (Invitrogen) by electroporation, spreading on LB-agar medium containing chloramphenicol (7 μg/ml) and ampicillin (4 μg/ml) and incubating at 25 °C for 42 hours. This very low-level ampicillin selection (below the MIC of the plasmid-free strain) reduces the frequency of improper constructs (typically deletion mutants) without eliminating variants that may serve as progenitors of the reference sequence. The thousands of colonies that grew were washed from the agar surface and grown in liquid culture with chloramphenicol (20 μg/ml; no ampicillin). Plasmid DNA was prepared from this mixed culture.

In parallel with the above, a single round of insertion mutagenesis was performed on a template where the fourth region (Figure 6) had been deleted from the TEM-1 *bla* gene. The wild-type gene has a single PstI site, which is present in all constructs used here. The two plasmid libraries (one resulting from successive insertion mutagenesis at the first three regions and the other from insertion mutagenesis at the fourth) were combined by cleavage and ligation at this PstI site. The result is a population of plasmids carrying a mixture of substitutions at all four regions, covering 49 amino acid positions throughout the large domain. This population was used to transform the TOP10 strain by electroporation, spreading on LB-agar with chloramphenicol (20 μg/ml) and ampicillin (5 μg/ml), and incubating at 25 °C for 42 hours. The resulting colonies (thousands) were washed from the agar surface and grown in liquid culture with chloramphenicol (20 μg/ml; no ampicillin). Substantially resistant clones were isolated from this culture by spreading on LB-agar with 10 μg/ml of ampicillin and incubating at 25 °C. Approximate ampicillin MIC values were determined for several clones that passed this selection. A clone showing better-than-average resistance (growing well at an ampicillin concentration of 40 μg/ml) was chosen as the progenitor of the reference sequence. Production of the reference sequence from this progenitor coincided with local side-chain randomization at residue set 4, as described below.

Local side-chain randomization

Because the genetic code tends to group codons according to hydropathic character, signature-consistent randomization is largely achievable by designing primers with appropriate base mixtures. Using the conventional symbols for nucleotide combinations (R=A, G; Y=C, T; K=G, T; S=C, G; V=A, C, G; N=A, C, G, T), the standard codon specifications used are as follows: NTK for positions scored *b*, VRW for positions scored *l*, NCT or RST for positions scored *i* (NCT if proline is represented), NYK for positions scored *m*, VVW for positions scored *c*, and NNK for positions scored *x*.

Supplementation of these standard specifications is desirable in the following cases. Position 112 is scored *c* rather than *l* because four sequences in the alignment show tyrosine residues (Figure 6). Because the standard VVW specification omits tyrosine, an additional primer specifying a tyrosine codon (TAT) was synthesized and used in the set 2 experiment in proportion to the representation of Tyr in the genetic code (i.e. one part TAT to 18 parts VVW). At position 207, the NYK specification was similarly supplemented to include tyrosine (one part TAT to 16 parts NYK). And at position 210, the NTK specification was supplemented to include tryptophan (one part TGG to 16 parts NTK).

The NNK and VRW specifications unavoidably introduce unwanted codon possibilities. The NNK specification includes the TAG stop codon as one of 32 possibilities, and the VRW specification includes codons for serine and glycine along with those for the intended hydrophilic amino acids. Taking this into account, the calculated proportion of signature-consistent sequences is >70% for sets 1, 2, and 4 (Table 2, column 8). These sets are handled by making appropriate adjustments to the estimated number of sequences tested (see below). In order to achieve a similarly high proportion of signature-consistent sequences for set 3, VAK specifications were used instead of VRW at positions 164 and 179 (both scored *l*), with arginine included by supplementation. The full specification for set 3 thereby achieves 91% signature compliance (Table 2, column 8). As discussed in the text, set 3 includes two positions, 161 and 163, that show strongly coupled conservation. An exception to the above codon-specification rules at position 161 enables one of the favored side-chains in this pair, Arg161, to be better represented than compliance with the hydropathic score (*x*; Figure 6) requires. Instead of using the NNK specification at this position (encoding Arg with 9% frequency), a VRW specification is used (encoding Arg with 25% frequency).

The reference sequence was obtained from the progenitor clone (see above) by performing local side-chain randomization at residue set 4. The progenitor plasmid was modified by replacing codon positions 193 through 214 in the variant *bla* gene with a StuI restriction site. After digesting this template plasmid with StuI, mixed-base primers incorporating the described codon specifications were used to fill in the missing genetic material by PCR. Gel-purified amplification products were ligated and used to transform the TOP10 strain by electroporation, spreading cells onto large trays containing LB-agar with chloramphenicol (20 μg/ml; no ampicillin). Various dilutions of the transformation culture were also spread on plates containing the same medium in order to estimate the total number of chloramphenicol-resistant clones (Table 2, column 1). Wrapped trays and plates were incubated at 37 °C for 20 hours. Colonies (numbering in the hundreds of thousands) were washed from the trays and thoroughly mixed. A 40 μl portion of mixture was used to inoculate 2 ml of 2× TY medium with chloramphenicol (20 μg/ml) for growth at 37 °C for eight hours in a rotary shaker. Cells from the resulting dense culture were subjected to ampicillin selection (10 μg/ml) by the quantitative protocol described above. Ampicillin plates and control plates were wrapped and incubated at 25 °C for 42 hours, after which colonies were counted on both (Table 2, columns 2 and 10).

One of the variants conferring ampicillin resistance was chosen as the reference sequence as described in the text (Results and Discussion, subsection Finding a reference sequence). Plasmid templates for local randomization at residue sets 1, 2, and 3 (Figure 6) were prepared by replacing the respective coding regions in the reference sequence with restriction sites, as was done for set 4. The three randomization experiments were then performed in parallel using the method described for set 4.

In each of the four randomization experiments, 30 colonies from the control plates were used for preparation of plasmid DNA, which was examined in order to measure the proportion of plasmids carrying a proper gene construct. It is typical in experiments of this kind for deletions of various sizes, often at the point of ligation (the junction), to reduce the throughput of proper constructs. For rapid assessment of junctions, each pair of mutational primers was designed (without altering the encoded sequence) to form a restriction site upon ligation. Absence of this site therefore signifies a junction defect. Results of restriction tests are shown in Table 2, column 6. DNA sequence analysis, performed on several plasmids that passed the restriction test, provides a measure of the frequency of fully proper gene constructs among plasmids that passed the restriction test (Table 2, column 7). Along with the calculated frequency of signature compliance among proper constructs (Table 2, column 8), discussed above, the frequencies in columns 6 and 7 enable estimation of the number of clones carrying proper signature-consistent constructs that were spread on ampicillin test plates in each experiment (Table 2, column 9). The desired pass rates are obtained from the ratio of the numbers in columns 10 and 9, as indicated (Table 2, footnote k).

Isolation of functional set 3 variants

The four randomized variant cell cultures from the above experiments were stored in aliquots as frozen stocks in liquid nitrogen. Since no ampicillin-resistant colony was isolated from the set 2 or set 3 experiments, frozen aliquots were thawed and used to inoculate 2 ml of 2× TY medium with chloramphenicol (20 μg/ml). These cultures, grown to high density at 37 °C, were diluted sixfold in LB medium and spread on plates containing 7 μg/ml of chloramphenicol and 10 μg/ml of ampicillin (35 μl per plate). Plates were incubated at 25 °C for 42 hours. The culture from the set 3 experiment produced several colonies, some of which were found to carry identical *bla* variants. Three distinct variants were identified (Figure 8). None was found from the set 2 experiment.

Acknowledgements

Sincere thanks to D. Alexander and T. Smith for general support, to L. LoConte, D. Williams, M. Mohan, and M. Stevens for helpful discussions, and to M. Mohan for help with proteome pattern searches.

Supplementary data

Supplementary data associated with this article, comprising details of the calculations, can be found, in the online version, at doi:10.1016/j.jmb.2004.06.058

References

1. Davidson, A. R., Lumb, K. J. & Sauer, R. T. (1995). Cooperatively folded proteins in random sequence libraries. *Nature Struct. Biol.* **2**, 856–863.
2. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA*, **93**, 5590–5594 see also pp. 7157–7166.
3. Keefe, A. D. & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.
4. Yamouchi, A., Nakashima, T., Tokuriki, N., Hosokawa, M., Nogamai, H., Arioka, S. *et al*. (2002). Evolvability of random polypeptides through functional selection within a small library. *Protein Eng.* **15**, 619–626.
5. Hayashi, Y., Sakata, H., Makino, Y., Urabe, I. & Yomo, T. (2003). Can an arbitrary sequence evolve towards acquiring a biological function? *J. Mol. Evol.* **56**, 162–168.
6. Yockey, H. P. (1977). On the information content of cytochrome c. *J. Theoret. Biol.* **67**, 345–376.
7. Reidhaar-Olson, J. F. & Sauer, R. T. (1990). Functionally acceptable substitutions in two α-helical regions of λ repressor. *Proteins: Struct. Funct. Genet.* **7**, 306–316.
8. Axe, D. D. (2000). Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. *J. Mol. Biol.* **301**, 585–596.
9. Taylor, S. V., Walter, K. U., Kast, P. & Hilvert, D. (2001). Searching sequence space for protein catalysts. *Proc. Natl Acad. Sci. USA*, **98**, 10596–10601.
10. Davidson, A. R. & Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA*, **91**, 2146–2150.
11. Axe, D. D., Foster, N. W. & Fersht, A. R. (1998). A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry*, **37**, 7157–7166.
12. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure* (2nd edit). Garland: New York.
13. Mazzariol, A., Cornaglia, G. & Nikaido, H. (2000). Contributions of the AmpC β-lactamase and the AcrAB multidrug efflux system in intrinsic resistance of *Escherichia coli* K-12 to β-lactams. *Antimicrob. Agents Chemother.* **44**, 1387–1390.
14. Strynadka, N. C. J., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C. *et al*. (1992). Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature*, **359**, 700–705.
15. Minasov, G., Wang, X. & Shoichet, B. K. (2002). An ultrahigh resolution structure of TEM-1 β-lactamase suggests a role for Glu166 as the general base in acylation. *J. Am. Chem. Soc.* **124**, 5333–5340.
16. Raffanti, E. F. & King, J. C. (1974). Effect of pH on the stability of sodium ampicillin solutions. *Am. J. Hosp. Pharm.* **31**, 745–751.
17. Yanagawa, H., Yoshida, K., Torigoe, C., Park, J. S., Sato, K., Shirai, T. & Go, M. (1993). Protein anatomy: functional roles of barnase module. *J. Biol. Chem.* **268**, 5861–5865.
18. Trainer, S., Bouthors, A.-T., Maveyraud, L., Guillet, V., Sougakoff, W. & Samama, J.-P. (2000). The high resolution crystal structure for class A β-lactamase PER-1 reveals the bases for its increase in breadth of activity. *J. Biol. Chem.* **275**, 28075–28082.
19. Chothia, C. & Lesk, A. (1986). The relationship between the divergence of sequences and structures in proteins. *EMBO J.* **5**, 823–826.
20. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
21. Miranker, A. D. & Dobson, C. M. (1996). Collapse and cooperativity in protein folding. *Curr. Opin. Struct. Biol.* **6**, 31–42.
22. Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, **27**, 1664–1670.
23. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321–353.
24. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–14.
25. Leabman, M. K., Huang, C. C., DeYoung, J., Carlson, E. J., Taylor, T. R., de la Cruz, M. *et al*. (2003). Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl Acad. Sci. USA*, **100**, 5896–5901.
26. Chothia, C. & Finkelstein, A. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039.
27. Crippen, G. M. & Maiorov, V. N. (1995). How many protein folding motifs are there? *J. Mol. Biol.* **252**, 144–151.
28. Govindarajan, S., Recabarren, R. & Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins: Struct. Funct. Genet.* **35**, 408–414.
29. Przytycka, T., Aurora, R. & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Struct. Biol.* **6**, 672–682.
30. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl Acad. Sci. USA*, **89**, 8721–8725.
31. Rozinov, M. N. & Nolan, G. P. (1998). Evolution of peptides that modulate the spectral qualities of bound, small-molecule fluorophores. *Chem. Biol.* **5**, 713–728.
32. Tsuji, T., Kobayashi, K. & Yanagawa, H. (1999). Permutation of modules or secondary structure units creates proteins with basal enzymatic properties. *FEBS Letters*, **453**, 145–150.
33. Matsuura, T., Ernst, A. & Plückthun, A. (2002). Construction and characterization of secondary structure modules. *Protein Sci.* **11**, 2631–2643.
34. Blanco, F. J., Angrand, I. & Serrano, L. (1999). Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study. *J. Mol. Biol.* **285**, 741–753.
35. Walsh, S. T. R., Lee, A. L., DeGrado, W. F. & Wand, A. J. (2001). Dynamics of a de novo designed three-helix bundle protein studied by ^{15}N ^{13}C, and ^{2}H NMR relaxation methods. *Biochemistry*, **40**, 9560–9569.
36. Coulson, A. F. W. & Moult, J. (2002). A unifold, mesofold, and superfold model of protein fold use. *Proteins: Struct. Funct. Genet.* **46**, 61–71.
37. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.
38. Lau, K. F. & Dill, K. A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl Acad. Sci. USA*, **87**, 638–642.

39. Bogarad, L. D. & Deem, M. W. (1999). A hierarchical approach to protein molecular evolution. *Proc. Natl Acad. Sci. USA*, **96**, 2591–2595.
40. Das, S. & Smith, T. F. (2000). Identifying nature's protein Lego set. *Advan. Protein Chem.* **54**, 159–183.
41. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Struct. Biol.* **9**, 553–558.
42. Ma, L., Ishii, Y., Ishiguro, M., Matsuzawa, H. & Yamaguchi, K. (1998). Cloning and sequencing of the gene encoding Toho-2, a class A β-lactamase preferentially inhibited by tazobactum. *Antimicrob. Agents Chemother.* **42**, 1181–1186.
43. George, D. G., Barker, W. C. & Hunt, L. T. (1990). Mutation data matrix and its uses. *Methods Enzymol.* **183**, 333–351.
44. Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* **244**, 332–350.
45. Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domains. *J. Mol. Biol.* **278**, 457–479.
46. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
47. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.

Edited by J. Thornton

(Received 28 November 2003; received in revised form 2 May 2004; accepted 18 June 2004)